Preliminary Examination:
LINEAR MODELS

Answer all questions and show all work.
Q1 is 35 points; Q2 is 30 points, and Q3 is 35 points.

1. Assume each Y_i ($i = 1, \ldots, n$) can be modeled by the following linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i,$$

where $\boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \sigma^2 \mathbf{V}$, $\sigma^2 > 0$, and

$$\mathbf{V} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = (1-\rho)\mathbf{I} + \rho\mathbf{J};$$

here, $\mathbf{I}$ is an $n \times n$ identity matrix; $\mathbf{J}$ is an $n \times n$ matrix whose elements are all 1s.

The 'centered form' of the model can be written as

$$Y_i = \beta_0 + \beta_1(X_{i1} - \bar{X}_1) + \cdots + \beta_p(X_{ip} - \bar{X}_p) + \epsilon_i,$$

where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$; $j = 1, \ldots, p$.

Define $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$; $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_n)'$; $\mathbf{j}$ is an n-dimensional vector of 1s; $\alpha = \beta_0 + \beta_1\bar{X}_1 + \cdots + \beta_p\bar{X}_p$; $\mathbf{X}_c = \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\tilde{\mathbf{X}}$ with

$$\tilde{\mathbf{X}} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}.$$

a. Show that the following is equivalent to the 'centered' form of the model:

$$\mathbf{Y} = (\mathbf{j}, \ \mathbf{X}_c)\begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\varepsilon}.$$

b. Let $\mathbf{X} = (\,\mathbf{j},\ \ \mathbf{X}_c\,)$ and $\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix}$. Derive the generalized least squares (GLS) estimator for $\boldsymbol{\beta}$ in terms of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{V}$.

c. Show that

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \begin{pmatrix} bn & \mathbf{0}' \\ \mathbf{0} & a\mathbf{X}_c'\mathbf{X}_c \end{pmatrix}$$

where $a = 1/(1-\rho)$ and $b = 1/[1+(n-1)\rho]$. (*Hint:* $\mathbf{V}^{-1} = a(\mathbf{I} - b\rho\mathbf{J})$.)

d. Show that

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \begin{pmatrix} bn\bar{Y} \\ a\mathbf{X}_c'\mathbf{Y} \end{pmatrix}.$$

e. Show that the GLS for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = \begin{pmatrix} \bar{Y} \\ (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y} \end{pmatrix},$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

2. Consider the cell means ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

for $i = 1,2,3$ and $j = 1,2,\ldots,n$, where ϵ_{ij} are iid $N(0,\sigma^2)$. The restriction

$$\mu_3 = \mu_1 - \mu_2$$

is placed on the parameters. Define $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)'$.

a. Write this as a general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and express the restriction in the form of $\mathbf{A}'\boldsymbol{\beta} = \boldsymbol{\delta}$.

b. Find the restricted least squares estimator, $\hat{\boldsymbol{\beta}}_R$. Express this estimator in terms of the treatment means, $\bar{Y}_{i.}$ for $i = 1,2,3$.

c. Define

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

and let $\hat{\boldsymbol{\beta}}$ denote the *unrestricted* least squares estimators. How do $Q(\hat{\boldsymbol{\beta}})$ and $Q(\hat{\boldsymbol{\beta}}_R)$ compare and why?

d. Find $E[Q(\hat{\boldsymbol{\beta}})]$ and var$[Q(\hat{\boldsymbol{\beta}})]$ (under the model without the restriction).

e. Consider testing $H_0 : \mu_3 = \mu_1 - \mu_2$. Give the *F test statistic* and its distribution when H_0 is true, **and** explain how this distribution will change under the alternative hypothesis H_a: $\mu_3 - (\mu_1 - \mu_2) = \delta \neq 0$.

3. Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ with rank $r \leq p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{V})$, where $\mathbf{V}$ is known and nonsingular. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ denote an ordinary least squares estimator, and $(\mathbf{X}'\mathbf{X})^{-}$ denotes a generalized inverse of $\mathbf{X}'\mathbf{X}$. Define:

$$\hat{\sigma}^2 = (n-r)^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{Y},$$

where $\mathbf{P}_X$ is the projection matrix onto the column space of $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$. Suppose that $\boldsymbol{\lambda}$ is a p-dimensional vector and $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable.

a. Suppose that $\mathbf{V} = \sigma^2\mathbf{I}$. Derive the sampling distribution of $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$.

b. Suppose that $\mathbf{V}\mathbf{X} = \mathbf{X}\mathbf{Q}$ for some matrix $\mathbf{Q}$. Prove that $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ and $(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$ are independent.

c. Suppose that $\mathbf{V} = \sigma^2(\mathbf{I} + \mathbf{P}_X)$, for some $\sigma^2 > 0$. Define

$$T = \frac{\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}'\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}}}.$$

Find the constant k such that kT follows a t distribution. What is its degrees of freedom? Is this a central t distribution?

d. As in part (c), suppose that $\mathbf{V} = \sigma^2(\mathbf{I} + \mathbf{P}_X)$. Use the results in part (c) to derive a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

e. As in part (c), suppose that $\mathbf{V} = \sigma^2(\mathbf{I} + \mathbf{P}_X)$. Build a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\lambda}'\boldsymbol{\beta}$, based on the generalized least squares estimator and compare the two intervals in (d) and (e).